D528–D532 *Nucleic Acids Research, 2004, Vol. 32, Database issue*
DOI: 10.1093/nar/gkh005

SNP500Cancer: a public resource for sequence validation and assay development for genetic variation in candidate genes

Bernice R. Packer*, Meredith Yeager, Brian Staats, Robert Welch, Andrew Crenshaw, Maureen Kiley, Andrew Eckert, Michael Beerman, Edward Miller, Andrew Bergen[1], Nathaniel Rothman[1], Robert Strausberg[2] and Stephen J. Chanock[3]

Intramural Research Support Program, SAIC-Frederick, NCI-FCRDC, Frederick, MD, USA, [1]Division of Cancer Epidemiology and Genetics, [2]Office of Cancer Genomics, National Cancer Institute, Bethesda, MD, USA and [3]Section on Genomic Variation, Pediatric Oncology Branch, National Cancer Institute, National Institutes of Health, Gaithersburg, MD, USA

Received July 8, 2003; Revised and Accepted August 7, 2003

ABSTRACT

The SNP500Cancer Database provides sequence and genotype assay information for candidate single nucleotide polymorphisms (SNPs) useful in mapping complex diseases, such as cancer. The database is an integral component of the NCI's Cancer Genome Anatomy Project. SNP500Cancer provides bi-directional sequencing information on a set of control DNA samples derived from anonymized subjects (102 Coriell samples representing four self-described ethnic groups: African/African-American, Caucasian, Hispanic and Pacific Rim). All SNPs are chosen from public databases and reports, and the choice of genes includes a bias towards non-synonymous and promoter SNPs in genes that have been implicated in one or more cancers. The web site is searchable by gene, chromosome, gene ontology pathway and by known dbSNP ID. As of July 2003, the database contains over 3400 SNPs, 2490 of which have been sequenced in the SNP500Cancer population. For each analyzed SNP, gene location and over 200 bp of surrounding annotated sequence (including nearby SNPs) are provided, with frequency information in total and per subpopulation, and calculation of Hardy–Weinberg Equilibrium (HWE) for each subpopulation. Sequence validated SNPs with minor allele frequency > 5% are entered into a high-throughput pipeline for genotyping analysis to determine concordance for the same 102 samples. The website provides the conditions for validated genotyping assays. SNP500Cancer provides an invaluable resource for investigators to select SNPs for analysis, design genotyping assays using validated sequence data, choose selected assays already validated on one or more genotyping platforms, and select reference standards for genotyping assays. The SNP500Cancer Database is freely accessible via the web page at http://snp500cancer.nci.nih.gov/.

INTRODUCTION

NCI and CGAP

SNP500Cancer is part of the National Cancer Institute's Cancer Genome Anatomy Project (CGAP) and is specifically designed to generate resources for the identification and characterization of genetic variation in genes important in cancer. CGAP (1) is dedicated to the development of technology, including both assays and utilization of technical platforms, to determine the gene expression profiles of normal, precancer and cancer cells. Accordingly, data pertaining to genes and their variation are made available on the public web site http://cgap.nci.nih.gov/. SNP500Cancer represents one of several initiatives designed to characterize sequence variation and is a resource for applying genetic approaches to understanding the etiology of different cancers as well as related phenotypes. Single nucleotide polymorphisms (SNPs) validated in this initiative are used by the NCI's Core Genotyping Facility (CGF) to genotype samples for studies coordinated by the Division of Cancer Epidemiology and Genetics (DCEG), the primary focus within the NCI for population-based research on environmental and genetic determinants of cancer.

Coriell samples

The SNP500Cancer initiative studies the genomes of 102 individuals of self-described heritage. The SNP500Cancer population is defined here as the sample of $n = 102$ DNAs with geographic origin and self-described ethnic group affiliation information to represent a diverse group of human populations. The anonymized samples are obtained from the Coriell Cell Repositories (Coriell Institute for Medical Research, Camden, NJ, USA), and represent four ethnic groups: 24

*To whom correspondence should be addressed. Tel: +1 301 496 6019; Fax: +1 301 402 3134; Email: packerb@mail.nih.gov

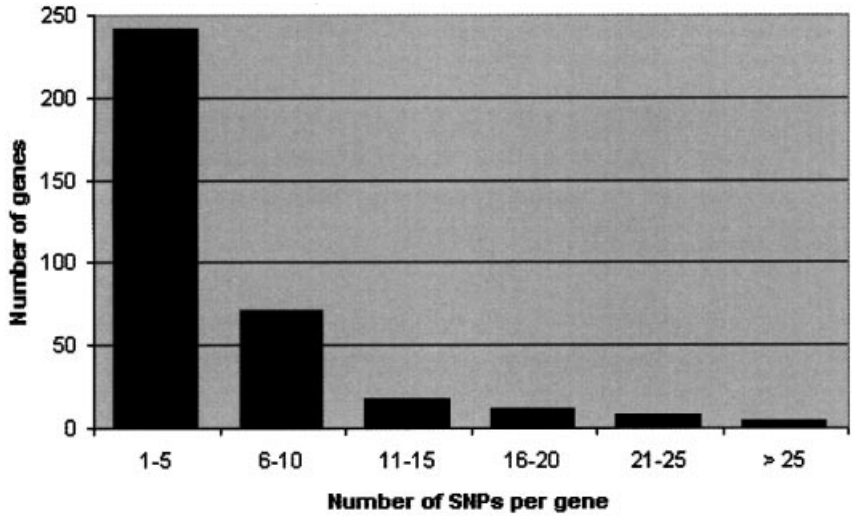


Figure 1. SNPs per gene.

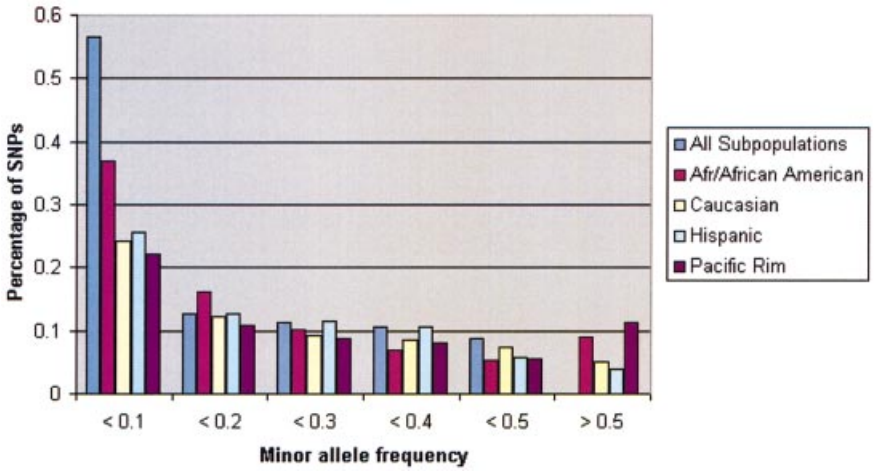


Figure 2. SNP500Cancer allele frequencies by subpopulation.

African/African-American, 31 Caucasian, 23 Hispanic and 24 Pacific Rim. These individuals are not a random sample of any specific human population, and thus the predictive value of the sequence and genotype data provided will vary for different population samples. However, where literature data are sparse, the allele frequencies in the SNP500Cancer population of n = 102 should provide assistance in determining how informative a given SNP is overall, as well as in each of the four subpopulations. It should also be noted that the SNP500Cancer subpopulations consist of subjects originally obtained from different geographic and ethnic groups. This approach was chosen for the purposes of discovery and validation of SNPs of interest to molecular epidemiology studies in cancer.

Selection of genes and SNPs

SNPs are chosen to be within or closely situated to candidate genes. The selection of genes and SNPs for analysis has been drawn from the following sources: (i) review of the published literature on SNPs and cancer, (ii) genes that fit a plausible model for cancer studies (e.g. by pathway), and (iii) SNPs reported in public databases with some associated non-*in silico* determined frequency.

As of July 2003, the database contains 480 genes. Figure 1 shows the distribution of number of validated SNPs per gene. The range is from 1 to 44 SNPs per gene, average = 5.4 SNPs per gene, median = 4 SNPs per gene.

Sequencing protocol

A contig of approximately 600 bp in length is generated for each SNP, which is localized to the center, creating flanking regions of roughly 300 bp in each direction. Additional putative SNPs (determined from dbSNP) are annotated onto the contig. Sequencing primers are designed for bi-directional sequence analysis using Primer3 software (2). Each primer is tagged with a universal sequencing primer, M13 (TGTAAAACGACGGCCAGT) for forward and M13

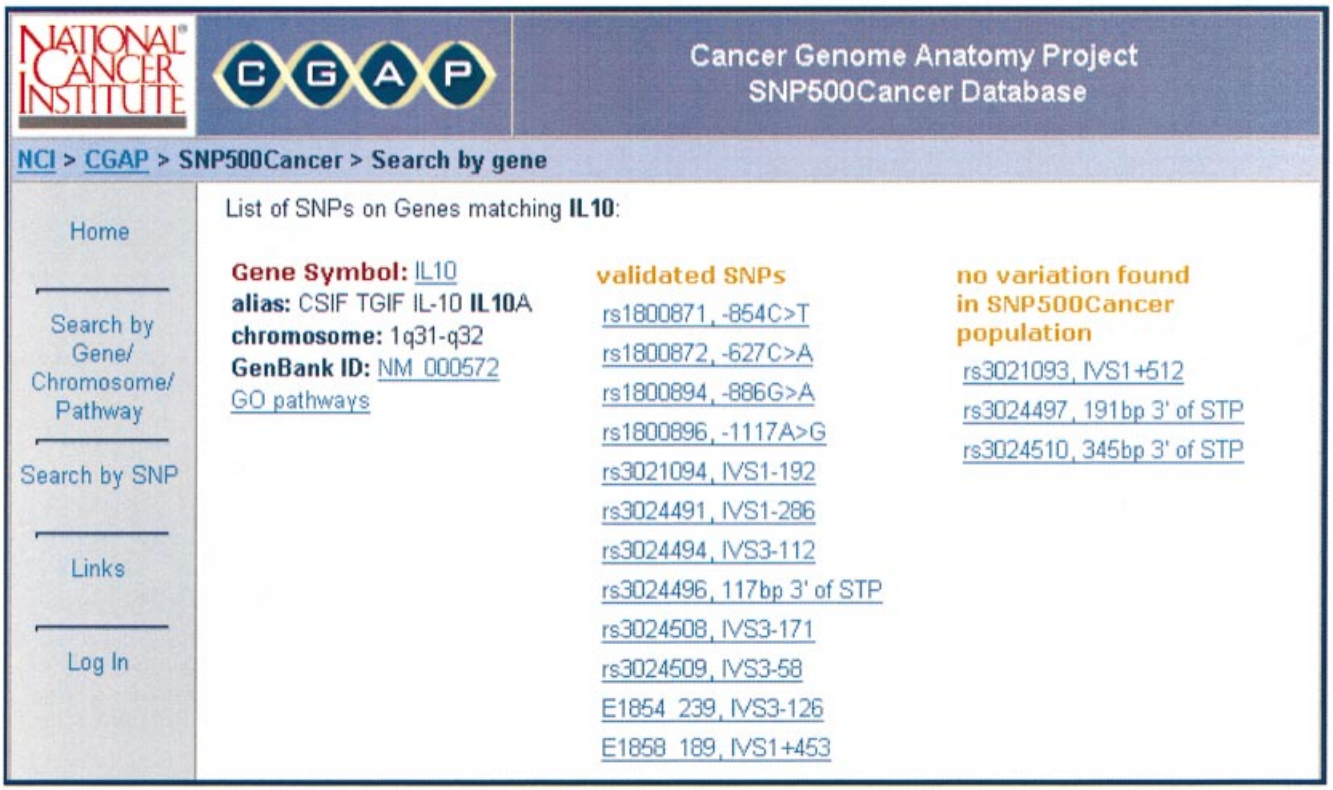


Figure 3. Listing a gene's SNPs.

(CAGGAAACAGCTATGACC) for reverse. The sequencing assay procedures and conditions are displayed on the SNP500Cancer website. Sequence tracings are analyzed in Sequencher 4.0.5 program (Genecodes, Ann Arbor, MI). After alignment of bi-directional sequence reads to the pre-annotated 600 bp contig, two independent reviewers analyze each contig for annotated and novel SNPs. The criteria for completing sequence alignment of each contig include 190 separate sequence tracings at a minimum of 70% assembly parameters. Genotype calls are determined for each of the 102 individuals and genotype and allele frequencies are maintained in an Oracle database and displayed on the SNP500Cancer website.

Genotyping protocols and validation

For SNPs that are determined to have >0.05 minor allele frequency in at least one of the SNP500Cancer subpopulations, approximately 200 bp of DNA sequence surrounding each SNP is submitted for design on one or more of the CGF's genotyping platforms: (i) Applied Biosystems' TaqMan™ 'Assay by Design' service, (ii) EPOCH Biosciences' MGB Eclipse™ probes, (iii) Sequenom Mass Array™. The genotyping assay procedures and conditions for all three platforms are displayed on the SNP500Cancer web site.

Genotypes are validated to establish concordance on two or more molecular genetic analysis platforms, where the primary comparison is between genotyping results from sequencing and from another genotyping platform, e.g. AB TaqMan™, Epoch MGB Eclipse™ or Sequenom MassARRAY™. A genotyping assay is validated when genotype analysis of the n = 102 DNA samples for that assay are concordant with genotypes determined from sequencing.

Allele frequencies

For each validated SNP, allele and genotype frequencies are displayed for the total SNP500Cancer population and for each SNP500Cancer subpopulation. For each analyzed SNP, a test for Hardy–Weinberg Equilibrium (HWE), χ^2 with one degree of freedom for two alleles (3) is performed per subpopulation. Figure 2 shows the distribution of minor allele frequencies in the four SNP500Cancer subpopulations.

dbSNP submission

All analyzed SNPs from the SNP500Cancer Database are submitted to dbSNP (4) http://www.ncbi.nlm.nih.gov/SNP/. This information includes flanking sequence, observed variation, assay primers, probes, and conditions, and frequency of the sequence variation among the SNP500Cancer total population and subpopulations.

USING THE SNP500CANCER WEBSITE

Searching for genes

The SNP500Cancer website provides capabilities for searching for genes in several ways: (i) gene name or alias (including wild card searches), (ii) chromosome location, (iii) Gene

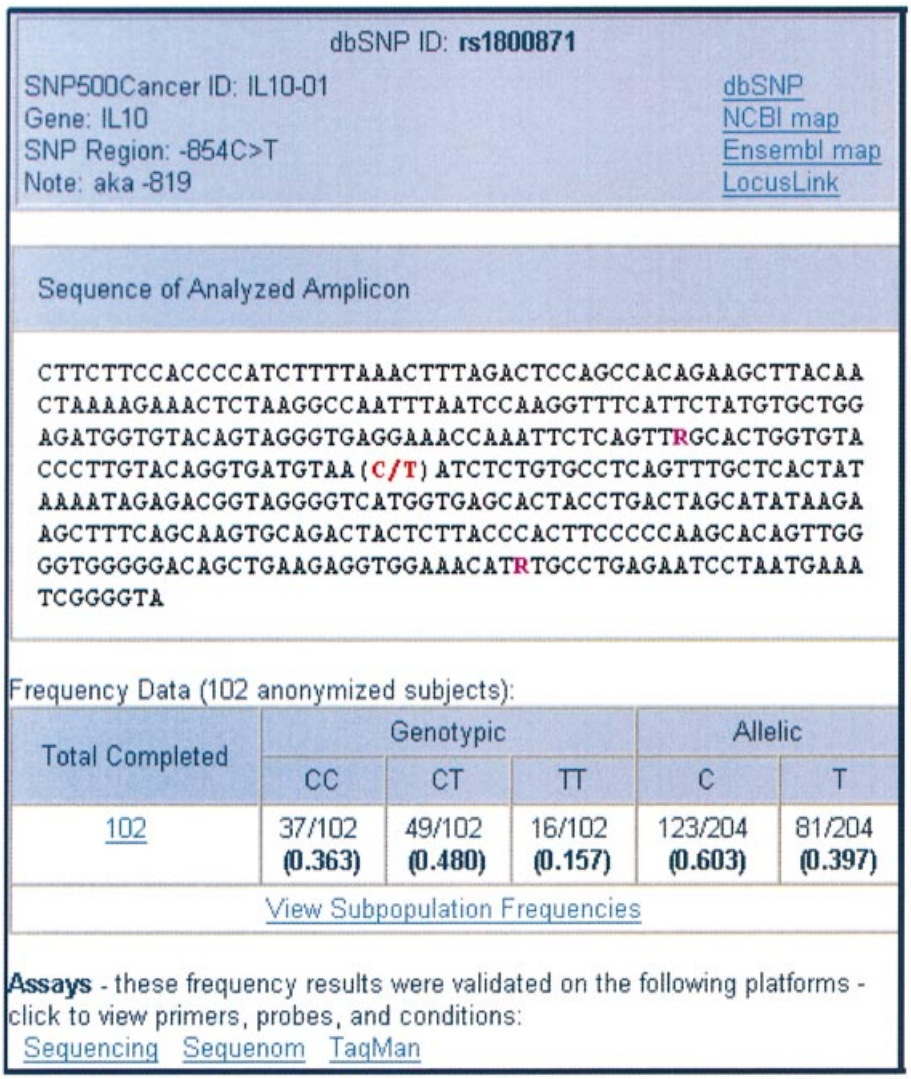


Figure 4. SNP information.

Ontology (GO) pathway (5)—numeric or text. The gene is displayed with a list of SNPs that have been validated, and those that were not found to occur in the SNP500Cancer population (Fig. 3).

Searching for SNPs

SNPs can be searched for using the dbSNP ID (rs cluster number, e.g. rs799917), or the internal SNP500Cancer polymorphism ID (gene symbol followed by a sequence number, e.g. BRCA1-02). The SNP is displayed in the center of surrounding sequence, and other SNPs in the sequence are annotated with IUPAC codes (Fig. 4).

Viewing genotypic and allelic frequencies

For the SNP of interest, genotypic and allelic frequencies for the entire SNP500Cancer population of 102 individuals are displayed. The 'View Subpopulation Frequencies' link displays a page with genotypic and allelic frequencies for each subpopulation—African/African-American, Caucasian, Hispanic, and Pacific Rim. Each subpopulation link leads to a list of individual genotypes for the samples within that subpopulation (Fig. 5).

Displaying assay conditions

For the SNP of interest, links are displayed for all validated assays (sequencing, MGB Eclipse™, Sequenom™, TaqMan™). Each link displays a page with detailed information on primers, probes, temperature and procedural steps.

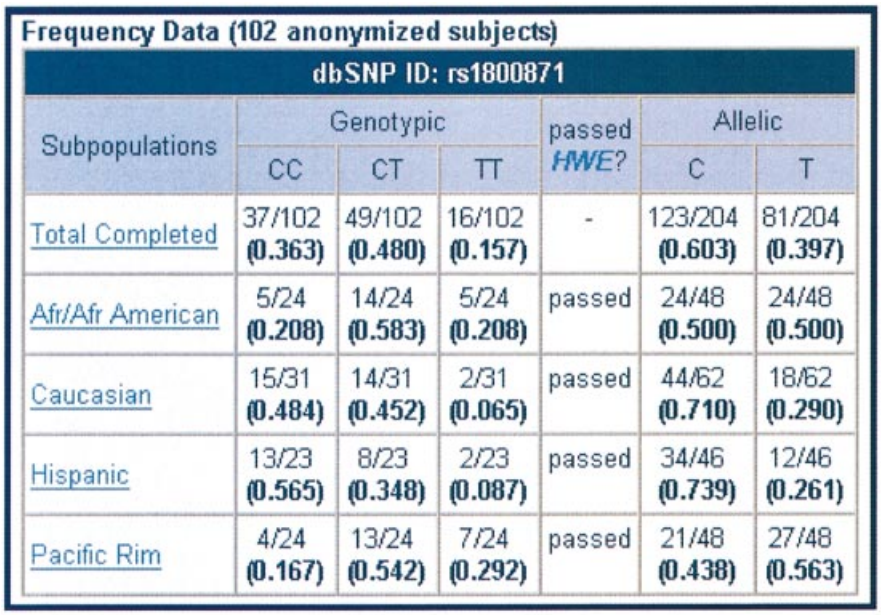
Connecting to other information sources

Each gene and SNP page on the SNP500Cancer website includes links to external resources. For genes: LocusLink (4), GO Database (5); for SNPs: dbSNP (4), NCBI MapViewer (4), Ensembl (6).

FUTURE DIRECTIONS

Additional genes and SNPs

As new potential associations between genes and diseases are investigated, the genes will be added to the SNP500Cancer



Frequency Data (102 anonymized subjects)

dbSNP ID: rs1800871

Subpopulations	Genotypic			passed HWE?	Allelic	
	CC	CT	TT		C	T
Total Completed	37/102 **(0.363)**	49/102 **(0.480)**	16/102 **(0.157)**	-	123/204 **(0.603)**	81/204 **(0.397)**
Afr/Afr American	5/24 **(0.208)**	14/24 **(0.583)**	5/24 **(0.208)**	passed	24/48 **(0.500)**	24/48 **(0.500)**
Caucasian	15/31 **(0.484)**	14/31 **(0.452)**	2/31 **(0.065)**	passed	44/62 **(0.710)**	18/62 **(0.290)**
Hispanic	13/23 **(0.565)**	8/23 **(0.348)**	2/23 **(0.087)**	passed	34/46 **(0.739)**	12/46 **(0.261)**
Pacific Rim	4/24 **(0.167)**	13/24 **(0.542)**	7/24 **(0.292)**	passed	21/48 **(0.438)**	27/48 **(0.563)**

Figure 5. Genotypic and allelic frequencies for a SNP.

Database. Additional SNPs will be selected based on their location within the genes, and citations in the published literature. This will increase the coverage across genes of interest.

Gene annotation server

The gene annotation server, currently under development, will serve as an information resource for gene and SNP annotations in the human genome comprised of public and CGF generated data. Users will be able to see SNPs in their genomic and proteomic contexts along with population-specific information, enabling them to make informative decisions on which SNPs to send through the CGF validation and assay pipelines. Data administrators will be able to create annotations based on CGF data directly through the user interface in an intuitive and productive manner. The annotations will be available to the public on the SNP500Cancer website, and will become an invaluable resource for investigators of gene-specific variations and their associations with disease.

Haplotype frequencies and htSNPs

All genes in the SNP500Cancer Database will be analyzed for inferred haplotypes occurring in the four subpopulations. Haplotypes will be presented graphically on the SNP500Cancer website. Haplotype tagging SNPs (htSNPs) will be calculated using different analytical methods. The htSNP information will allow fewer SNPs to be genotyped per gene, thereby reducing cost and improving throughput.

REFERENCES

1. Strausberg,R.L., Simpson,A.J.G. and Wooster,R. (2003) Sequence-based cancer genomics: progress, lessons, and opportunities. *Nature Rev. Genet.*, **4**, 409–418.
2. Rozen,S. and Skaletsky,H.J. (2000) Primer3 on the WWW for general users and for biologist programmers. *Methods Mol. Biol.*, **132**, 365–386.
3. Weir,B.S. (1996) *Genetic Data Analysis II: Methods for Discrete Population Genetics Data*. Sinauer, Sunderland, MA.
4. Wheeler,D.L., Church,D.M., Federhen,S., Lash,A.E., Madden,T.L., Pontius,J.U., Schuler,G.D., Schrim,L.M., Sequeira,E., Tatusova,T.A. and Wagner,L. (2003) Database resources of the National Center for Biotechnology. *Nucleic Acids Res.*, **31**, 28–33.
5. Gene Ontology Consortium. (2001) Creating the gene ontology resource: design and implementation. *Genome Res.*, **11**, 1425–1433.
6. Hubbard,T., Barker,D., Birney,E., Cameron,G., Chen,Y., Clark,L., Cox,T., Cuff,J., Curwen,V., Down,T. *et al.* (2002) The Ensembl genome database project. *Nucleic Acids Res.*, **30**, 38–41.